

Integrated Cellular and *Ad Hoc* Relaying Systems: iCAR

Hongyi Wu, Chunming Qiao, Swades De, and Ozan Tonguz

Abstract—Integrated cellular and *ad hoc* relaying systems (iCAR) is a new wireless system architecture based on the integration of cellular and modern *ad hoc* relaying technologies. It addresses the congestion problem due to unbalanced traffic in a cellular system and provides interoperability for heterogeneous networks. The iCAR system can efficiently balance traffic loads between cells by using *ad hoc* relaying stations (ARS) to relay traffic from one cell to another dynamically. This not only increases the system's capacity cost effectively, but also reduces transmission power for mobile hosts and extends system coverage. In this paper, we compare the performance of the iCAR system with conventional cellular systems in terms of the call blocking/dropping probability, throughput, and signaling overhead via analysis and simulation. Our results show that with a limited number of ARSs and some increase in the signaling overhead (as well as hardware complexity), the call blocking/dropping probability in a congested cell and the overall system can be reduced.

Index Terms—*Ad hoc*, blocking probability, cellular, load balancing, mobile, relaying, wireless.

I. INTRODUCTION

TRADITIONAL cellular systems have provided voice services since the first analog system was introduced about 15 years ago. In the last decade, with the unprecedented increase in demand for personal mobility and dependence on personal communications, both the number of subscribers and the amount of wireless traffic have surged at an exploding speed. With the advent of the Internet, especially the wireless access to the Internet, wireless data traffic is expected to exacerbate the demand for bandwidth. The carriers and infrastructure providers now face a major challenge in meeting the increased bandwidth demand of mobile Internet users.

At the same time, efforts in providing various access services such as wireless LANs, *ad hoc* networks, Bluetooth, and home RF networks, are further stimulating the growth of wireless traffic and the requirement for an ubiquitous wireless infra-

structure. More specifically, continued proliferation of these services will call for interoperability between heterogeneous networks such as *ad hoc* and cellular systems. In addition, such an interoperability will create even heavier traffic in cellular systems as more and more traffic from wireless LANs, *ad hoc* networks, and Bluetooth devices will be carried by the cellular infrastructure.

For the reasons cited above and the fact that the traffic in future cellular systems will be more bursty and unevenly distributed than conventional voice traffic, it is anticipated that *congestion* will occur in peak usage hours even in the next generation [e.g., third generation (3G)] systems, despite its increased capacity. By congestion, we mean that in some cells, data channels (DCHs) are less frequently available than the minimum acceptable level and as a result, the grade of service (GoS) in those cells has deteriorated below a prescribed threshold level (e.g., the call blocking probability in those cells becomes higher than 2%). Note that, however, control channels (CCHs) for signaling (or paging) *may* still be accessible by all mobile hosts (MHs) in a congested cell.

The presence of *unbalanced traffic* will exacerbate the problem of limited capacity in existing wireless systems. In a cellular system, an MH can use only the data channels of the base transceiver station (BTS) located in the same cell, which is a subset of the data channels available in the system. No access to data channels in other cells by the MH limits the channel efficiency and consequently the system capacity. Specifically, some cells may be heavily congested (called *hot spots*), while the other cells may still have enough available DCHs. In other words, even though the traffic load does not reach the maximum capacity of the entire system, a significant number of calls may be blocked and dropped due to localized congestion. Since the locations of hot spots vary from time to time (e.g., downtown areas on Monday morning, or amusement parks on Sunday afternoon), it is difficult, if not impossible, to provide the guarantee of sufficient resources in each cell in a cost-effective way. In fact, increasing the bandwidth of a cellular system (e.g., the number of DCHs in each cell) can increase the system capacity but not the efficiency to deal with the time-varying unbalanced traffic.

In this work, we address the important problem of how to evolve from the existing heavily invested cellular infrastructure to next generation wireless systems that scale well with the number of mobile hosts and, in particular, overcome the congestion by dynamically balancing the load among different cells in a cost-effective way. The basic idea of the proposed system called iCAR is to place a number of *ad hoc* relaying stations (ARSs) at strategic locations, which can be used to relay signals

Manuscript received December 1, 2000; revised June 1, 2001. This material is based upon work supported by the National Science Foundation under Grant 0082916.

H. Wu is with the Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY 14260 USA (e-mail: hongyiwu@cse.buffalo.edu).

C. Qiao is with the Department of Computer Science and Engineering and the Department of Electrical Engineering, State University of New York at Buffalo, Buffalo, NY 14260 USA (e-mail: qiao@computer.org).

S. De is with the State University of New York at Buffalo, Buffalo, NY 14260 USA. (e-mail: swadesd@eng.buffalo.edu).

O. Tonguz was with the Electrical Engineering Department of the State University of New York at Buffalo, Buffalo, NY 14260 USA. He is now with the Electrical and Computer Engineering Department, Carnegie Mellon University, Pittsburgh, PA 15213-3890 USA (e-mail: tonguz@ece.cmu.edu).

Publisher Item Identifier S 0733-8716(01)08475-X.

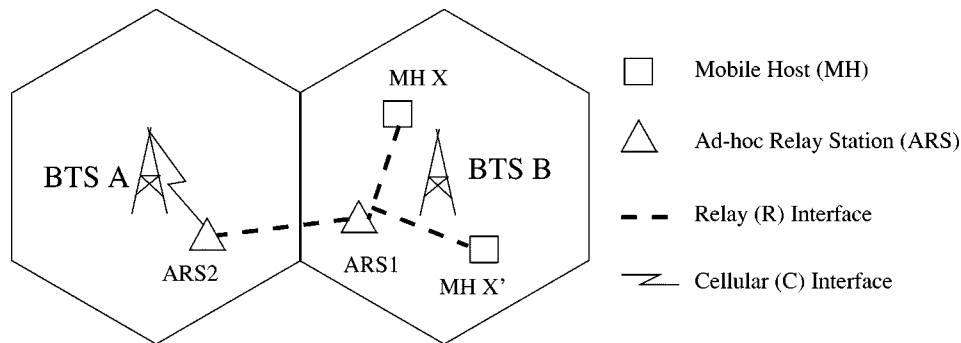


Fig. 1. A relaying example where MH X communicates with BTS through two *ad hoc* relaying stations (ARSs) (it may also communicate with MH X' through ARS 1).

between MHs and BTSs [1], [2]. By using ARSs, it is possible to divert traffic from one (possibly congested) cell to another (non-congested) cell. This helps to circumvent congestion and makes it possible to *maintain* (or hand-off) calls involving MHs that are moving into a congested cell, or to accept new call requests involving MHs that are in a congested cell. Although we will only focus on the issues related to load balancing in this paper, there are many other benefits of the proposed iCAR system. For example, the ARSs can, in a flexible manner, extend cellular system's coverage (similar to the wireless routers used in the Rooftop system[3]) and provide interoperability between heterogeneous systems (by connecting *ad hoc* networks and wireless LANs to the Internet for example). Additional benefits include enhanced reliability (or fault-tolerance) of the system and potential improvement in MHs' battery life and transmission rate.

In this paper, we evaluate the performance of the iCAR system via analysis and simulation. The predictions of our analysis are verified by simulation results, which are obtained using a more realistic model than the one used in [1] and [2] that simulated only static traffic. The call dropping/blocking probability, throughput, and additional signaling overhead introduced by relaying are the main metrics used for evaluating the performance of the proposed iCAR system. Our results indicate that with a limited number of ARSs, an iCAR system is able to efficiently balance the traffic load among cells which, in turn, leads to significantly lower call blocking and dropping probabilities than that in a corresponding cellular system.

The remainder of this paper is organized as follows. Section II reviews the principle of operation and main benefits of the proposed iCAR system. Section III presents the analysis of the iCAR system performance. Section IV evaluates the performance of the iCAR system through simulations and compare the proposed iCAR system with a conventional cellular system without load balancing, in terms of call blocking/dropping probability, throughput, and overhead in congested cells as well as the overall systems. Section V discusses related work in the literature. Finally, Section VI concludes the paper.

II. AN OVERVIEW OF THE iCAR SYSTEM

In this section, we describe the principle of operation and the main benefits of iCAR (see [2] for more details). To simplify the following presentation, we will focus on cellular systems where

each BTS is controlled by a mobile switching center (MSC) [4], [5] (although the concept also applies to radio network controller (RNC) in 3G systems). Major differences between BTSs and the proposed ARSs are as follows. Once a BTS is installed, its location is fixed since it often has a wired interface to an MSC (and a backbone network). An ARS, on the other hand, is a *wireless* communication device deployed by a network operator. It has much lower complexity and fewer functionalities than that needed for a BTS. In addition, it may, under the control of an MSC, have limited mobility (in order to adapt to varying traffic patterns)¹ and communicate *directly* with a BTS, another ARS, or an MH through the appropriate air interfaces.

An example of relaying is illustrated in Fig. 1, where MH X in cell B (congested) communicates with the BTS in cell A (or BTS A, which is noncongested) through two ARSs (there will be at least one ARS along which a *relaying route* is set up). Note that each ARS has two air interfaces, the **C** (for cellular) interface for communications with a BTS and the **R** (for relaying) interface for communicating with an MH or another ARS. Also, MHs should have two air interfaces; the C interface for communicating with a BTS and the R interface for communicating with an ARS. In the following discussion, we will assume that the C interface operates at or around 1900 MHz (PCS), and the R interface uses an unlicensed band at 2.4 GHz (in the ISM band), even though our concept also applies when different bands are used (for example, 850 MHz for the C interface as in 2G systems or 2 GHz for 3G systems). The R interface (as well as the medium access control (MAC) protocol used) is similar to that used in wireless LANs or *ad hoc* networks (see for example [6]–[15]). Note that because multiple ARSs can be used for relaying, the transmission range of each ARS using its R interface can be much shorter than that of a BTS, which implies that an ARS can be much smaller and less costly than a BTS. At the same time, it is possible for ARSs to communicate with each other and with BTSs at a higher data rate than MHs can, due to limited mobility of ARSs and specialized hardware (and power source).

There are three basic relaying strategies.

Primary Relaying: In an existing cellular system, if MH X is involved in a new call (as a caller or callee) but it is in a congested cell B, the new call will be blocked. In the proposed system with integrated cellular and relaying technologies, the

¹In this study, however, we only consider static ARSs. We intend to examine the benefit of ARSs with limited mobility in future work.

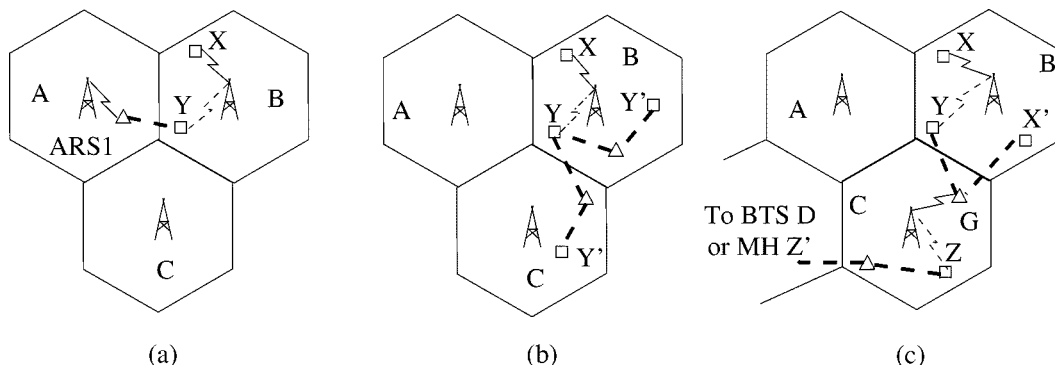


Fig. 2. Secondary relaying to free up a channel for MH X: (a) MH Y to BTS A, (b) MH Y to MH Y', or (c) cascaded relaying (i.e., MH Y to BTS C and MH Z to either MH Z' or BTS D).

call may not have to be blocked. More specifically, MH X which is in the congested cell B can *switch over* to the R interface to communicate with an ARS in cell A, possibly through other ARSs in cell B (see Fig. 1 for an example). We call this strategy *primary relaying*.

With primary relaying, MH X can communicate with BTS A, albeit indirectly (i.e., through relaying). Hereafter, we will refer to the process of changing from the C interface to the R interface (or vice versa) as switching-over, which is similar to (but different from) frequency hopping [4], [16], [17]. Of course, MH X may also be relayed to another nearby noncongested cell other than cell A. A relaying route between MH X and its corresponding (i.e., caller or callee) MH X' may also be established (in which case, both MHs need to switch over from their C interfaces to their R interfaces), even though the probability that this occurs is typically very low.

Secondary Relaying: If primary relaying is not possible, because, for example in Fig. 1, ARS 1 is not close enough to MH X to be a proxy (and there are no other nearby ARSs), then one may resort to *secondary relaying* so as to *free up* a DCH from BTS B for use by MH X. Two basic cases are illustrated in Fig. 2(a) and (b), respectively, where MH Y denotes any MH in cell B which is currently involved in a call. More specifically, as shown in Fig. 2(a), one may establish a relaying route between MH Y and BTS A (or any other cell). In this way, after MH Y switches over, the DCH used by MH Y can now be used by MH X. Similarly, as shown in Fig. 2(b), one may establish a relaying route between MH Y and its corresponding MH Y' in cell B or in cell C, depending on whether MH Y is involved in an intracell call or an intercell call. Note that congestion in cell B implies that there are a lot of on-going calls (involving candidates like MH Y); hence, the likelihood of secondary relaying [refer to Fig. 2(a) and (b)] should be better than that of primary relaying (refer to Fig. 1). In addition, although the concept of having an MH-to-MH call via ARSs only (i.e., no BTSs are involved) is similar to that in *ad hoc* networking, a distinct feature (and advantage) of the proposed integrated system is that an MSC can perform (or at least assist in performing) critical call management functions such as authentication, billing, and locating the two MHs and finding and/or establishing a relaying route between them, as mentioned earlier. Such a feature is also important to ensure that switching-over of the two MHs (this concept is not applicable to *ad hoc* networks) is completed fast

enough so as not to disconnect the on-going call involving the two MHs or not to cause severe quality of service (QoS) degradation (even though the two MHs may experience a “glitch” or jitter).

Cascaded Relaying: If neither primary relaying, nor basic secondary relaying [as shown in Fig. 2(a) and (b)] works, the new call may still be supported. More specifically, assume that there is a relaying route, which can be either primary or secondary relayed, between MH X and ARS, say G (for gateway), in a nearby cell C which unfortunately is *congested*. As shown in Fig. 2(c), one may apply any of the two basic secondary relaying strategies described above in the congested cell C (i.e., in a *cascaded* fashion) to establish a relaying route between an MH (say MH Z) in cell C and either another BTS in a noncongested cell or MH Z'. In this way, ARS G can be allocated the DCH previously used by MH Z in cell C, and, in turn, MH X can be allocated the DCH previously used by MH Y in cell B if the route between MH X and ARS G is set up by secondary relaying.

In addition to the above relaying strategies, one critical design issue in iCAR is the number and placement of ARSs. In [1], we have discussed the maximum number of relaying stations needed to ensure that a relaying route can be established between any BTS and an MH located anywhere in any cell. In the case where only a limited number of ARSs is available, an approach called *seed growing*, whereby one *seed ARS* is placed on each edge as shown in Fig. 5, can be used (note that additional ARSs may be placed around these seeds to increase the ARS coverage). Consequently, traffic in the ARS coverage area in one cell can be relayed to a neighboring cell covered by the same seed ARS (provided that it will not be blocked in that neighboring cell). It has been shown that, for an n -cell system, the maximum number of seed ARSs needed is $3n - \lfloor 4\sqrt{n} - 4 \rfloor$ [2]. In the following analysis and simulations, we assume that the seed growing approach is used and denote the ARS coverage in terms of the percentage of a cell covered by ARSs, by $0 < p \leq 1$.

III. PERFORMANCE ANALYSIS OF THE iCAR SYSTEM

In this section, we evaluate the performance of the iCAR system via analysis.

A. Principles

We first discuss the principle for the performance improvement of the iCAR system over a conventional cellular system assuming that the entire system can be covered by ARSs (i.e., $p = 1$) so that an MH in a cell can reach the BTS in any cell in the system via relaying. We present the following two theorems to show that iCAR will outperform the conventional cellular system. The first theorem states the best performance that a conventional cellular system can achieve.

Theorem 1: Assume that the total traffic in an n -cell system is T Erlangs, then the (system wide) call blocking probability is minimized when the traffic in each cell is T/n Erlangs.

The proof of this theorem is given in Appendix A. This theorem shows that as a result of being able to distribute traffic evenly in the system, the call blocking probability will be minimized.

Note that, unlike a conventional system where channel borrowing is limited by cellular band interference, an ideal n -cell iCAR system where an MH can be relayed to any BTS can be treated as a single *super* cell system with n times of DCHs. Given the same total traffic T Erlangs, the call blocking probability in the super cell is lower than that of a conventional cellular system even when the traffic is evenly distributed among the n cells. More formally, we have the following theorem.

Theorem 2: For a given total traffic in a system and a fixed number of DCHs in each cell, an ideal iCAR has a lower blocking probability than any conventional cellular systems (including a perfectly load-balanced one).

The proof of this theorem is in Appendix B. Note that the above two theorems serve as a proof of principle that iCAR can perform better than any conventional cellular systems. However, what has been implicitly assumed is that, in the ideal iCAR, not only are there sufficient numbers of ARSs, but also there is no bandwidth shortage along any relaying route such that any number of calls can be relayed through an ARS.

B. Analysis

In this section, we analyze the performance of iCAR with limited ARS coverage p using the Erlang-B model [4]. We partition an iCAR system with unbalanced traffic and scattered hot spots into subsystems. Each subsystem includes a hot spot at the center and the traffic in it is assumed to be location-dependent (i.e., the farther away from the hot spot, the lower the traffic intensity is).² Since there is little or no interaction (e.g., relaying) among cells in different subsystems, the analysis will focus only on a three tier subsystem shown in Fig. 3. More specifically, we denote the traffic intensity in cell A, each tier B cell, and each tier C cell in the absence of relaying by T_a , T_b , and T_c , respectively, and the corresponding call blocking probabilities by B_a , B_b , and B_c , respectively. If perfect load balancing is achieved, the traffic intensity per cell will be

$$T_f = \frac{T_a + 6T_b + 12T_c}{19}. \quad (1)$$

²Similar techniques can also be applied to subsystems with other traffic patterns.

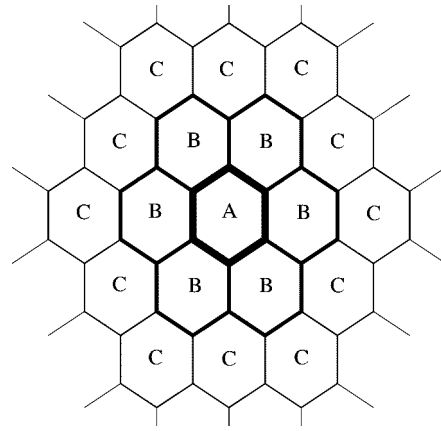


Fig. 3. A three tier subsystem considered in our analysis.

We further assume that in the absence of load balancing, cell A is a hot spot, which is surrounded by “cooler” tier B cells and even cooler tier C cells, as depicted in Fig. 3. In other words, we assume that $T_a > T_b > T_c$, and in addition, $T_a > T_f$, $T_c < T_f$, and T_b could be larger, equal, or smaller than T_f . The difference between the actual traffic load in each cell and the subsystem-wide average load, namely $T_i - T_f$ (where $i = a, b, c$), will determine the amount of load balancing desired (which may not be achievable due to limited ARS coverage and blocking in neighboring cells). Below, we provide a steady-state solution for the traffic intensities reached after achieving dynamic load balancing via primary and secondary relaying.

To facilitate our analysis, we assume that traffic is not spread from a cell to an equally loaded or a more heavily loaded cell, and in addition, traffic is evenly distributed in any given cell, and hence, the probability that a call can be relayed to a neighboring cell (provided that it will not be blocked in that neighboring cell) is equal to the fractional ARS coverage p . We do not consider cascaded relaying in our analysis, as in the three-tier model, cascaded relaying results in little or no improvement over secondary relaying.

1) *Primary Relaying:* Since primary relaying will attempt to transfer only the amount of overload traffic (which most likely represents blocked calls) to/from MHs covered by ARSs to cell Bs, assuming that the traffic in a cell is uniformly distributed, the average amount of overload traffic in cell A that can be transferred via primary relaying to tier B cells is $p(T_a - T_f)(1 - B_b)$. Hence, the average traffic load in cell A after primary relaying becomes

$$T_a^p = T_a - p(T_a - T_f)(1 - B_b). \quad (2)$$

Accordingly, the new call blocking probability in cell A due to primary relaying is

$$B_a^p = \frac{(T_a^p)^M / M!}{\sum_{i=0}^M (T_a^p)^i / i!} \triangleq f(T_a^p, M) \quad (3)$$

where M is the number of cellular band channels. Note that, although as a result of primary relaying the traffic load in the

ARS coverage area in cell A may have been reduced, the load in other areas in cell A (i.e., areas not covered by ARSs) has not, and the total load in cell A is higher than T_f . More specifically, the average amount of overload traffic in cell A becomes $U_a^p = T_a^p - T_f$, which is still nonnegative and can only be reduced via secondary relaying as to be discussed in the next subsection.

Since the average amount of overload traffic relayed from cell A to each of the six tier B cells is $(p/6)(T_a - T_f)(1 - B_b)$, the traffic load in each B cell becomes $T' = T_b + (p/6)(T_a - T_f)(1 - B_b)$. This, however, will be reduced due to primary relaying of traffic from tier B to tier C cells (which initially have a lower blocking probability than Bs). More specifically, since each B cell is surrounded by three C cells, traffic relayed from a cell B to tier C cell is $(p/2)(T' - T_f)(1 - B_c)$. Hence, the average traffic in cell B becomes

$$T_b^p = T' - \frac{p}{2}(T' - T_f)(1 - B_c). \quad (4)$$

Accordingly, the new call-blocking probability in cell B due to primary relaying is $B_b^p = f(T_b^p, M)$, which is obtained from (3) by replacing T_a^p with T_b^p .

Similarly, one can compute the average traffic in cell Cs after primary relaying and the corresponding new call blocking probability.

2) *Secondary Relaying*: The goal of secondary relaying is to distribute the load more evenly than what is possible via primary relaying. For cell A, this is accomplished by trying to relay additional traffic in the ARS coverage area (which most likely represents on-going calls) in order to offset (i.e., reduce) the higher than average traffic load in the entire cell A. Recall that the overload traffic in cell A after primary relaying is $U_a^p = T_a^p - T_f$, which is the excess amount that one *ideally* would like to transfer to tier B cells. However, based on the previous discussion, the traffic that can be transferred via secondary relaying in cell A is at most $pT_a^p(1 - B_a^p)(1 - B_b^p)$. Hence, the traffic that will be transferred via secondary relaying from cell A is

$$R_a^s = \min\{U_a^p, pT_a^p(1 - B_a^p)(1 - B_b^p)\}. \quad (5)$$

As a result of secondary relaying, the average traffic in cell A becomes

$$T_a^s = T_a^p - R_a^s \quad (6)$$

based on which, the corresponding new call blocking probability in cell A becomes $B_a^s = f(T_a^s, M)$.

Similarly, one can compute the adjusted traffic load in tier B and C cells and the new blocking probability after secondary relaying.

C. Analytical Results

Without loss of generality, we assume that each BTS has $M = 50$ DCHs and T_a is 50 *Erlangs* which corresponds to 5% blocking probability in cell A. We also assume that the traffic intensity decreases to 0.8 fraction from one tier of cells to another, which means that $T_b = 0.8T_a$ and $T_c = 0.8T_b$, and consequently results in approximately 1.87% and 0.75% blocking probability in tier B and C cells, respectively.

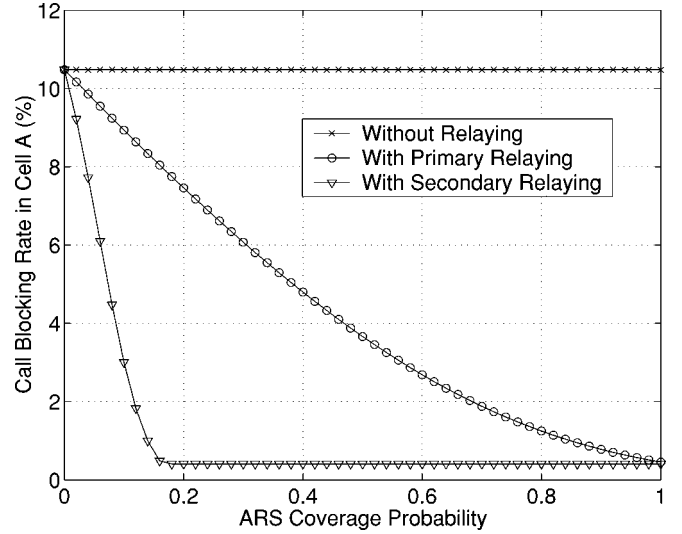


Fig. 4. Call blocking probability versus ARS coverage probability when $T_a = 50$ *Erlangs*.

Fig. 4 shows the impact of ARS coverage (p) on the call blocking probability in cell A, with primary relaying only and with secondary relaying.³ Observe that to achieve an acceptable call blocking probability (e.g., 2%) with primary relaying only, ARS coverage has to be very high compared to that with secondary relaying. This is because primary relaying is effective only on the blocked calls, whereas secondary relaying operates on the ongoing calls which are much larger in number compared to the blocked calls.

Additional analytical results will be presented along with the simulation results in the next section.

IV. SIMULATION RESULTS

To obtain performance results under more realistic assumptions, we have also developed a simulation model. As in the analysis, we partition the system with unbalanced traffic and scattered hot spots into subsystems. In this simulation, we study only one subsystem (see the area inside the dashed rectangle in Fig. 5), which is quite similar to the model used in the analysis (shown in Fig. 3) except for several additional cells (in tier D).

The average call arrival rate and holding time are two factors determining the traffic load (measured in *Erlangs*) in a cell. To facilitate our simulation of different traffic intensities, we keep the average call generation rate fixed and vary the average call holding time (note that we could have varied the call generation rate instead). The holding time is a random variable with cut negative exponential distribution. Table I(b) gives an example of mapping from average holding time to traffic intensities we get from the simulation.

There are $5 \times 5 = 25$ BTSs and 56 seed ARSs in the simulation model. We assume that the longest transmission range of a BTS is 2 Km and an ARS (which is placed at each shared border of two adjacent cells) covers an area whose radius is 500 m. This results in the ARS coverage of $p = 0.23$. Each BTS

³It should be noted that *secondary relaying*, by definition, is a relaying strategy that includes, as a first step, the use of primary relaying, or in other words, secondary relaying is implemented in addition to primary relaying.

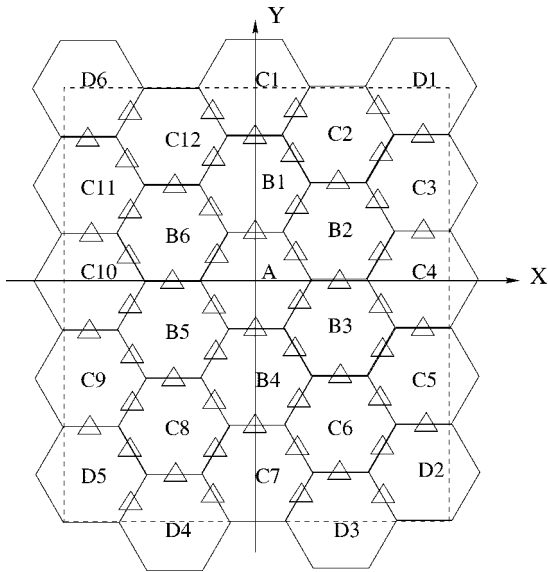


Fig. 5. Simulation environment.

has 50 cellular band channels (i.e., $M = 50$), and by default each ARS can handle up to three cellular band channels using a proper multiplexing technique. In order to obtain good statistical results, over 25 000 MHs are simulated which are initially placed in the system with uniform distribution. Table I(a) lists the parameters used in the simulation.

The simulations were performed using GloMoSim [18]. In addition to the operations in a conventional cellular system (including handoffs from one BTS to another), we implement primary, secondary, cascaded relaying, and various other handoffs (e.g., from a BTS to an ARS and from an ARS to a BTS). As mentioned in footnote 3, when we talk about the performance of secondary relaying, it implies that both primary and secondary relaying are implemented. Similarly, cascaded relaying actually includes primary and secondary relaying. The call dropping/blocking probability, throughput, and additional signaling overhead introduced by relaying are the main metrics used to evaluate the performance of both cell A and the entire subsystem.⁴ The *random waypoint model* wherein an MH selects a random speed, moves for 8 s, stays there for 2 s, and then starts to move again, is used to simulate different mobilities to study their effects on handoffs [19] and call dropping probabilities. The movement of MHs is limited within the dashed square area (which only has a few additional cell Ds to simplify the simulation model). The moving direction is random from 0° to 360° . The absolute speed value is a random number within a range between 0 meter per second (m/s) and a specified maximum speed. In order to obtain converged results, we run the simulation for 10 h for each traffic intensity and MH mobility combination before collecting the results. The MHs in the system generate over 250 000 calls during this period.

A. Call Blocking Probability

A new call is blocked if there is no free DCH available when it is generated. Fig. 6 shows the results for call-blocking proba-

⁴Call blocking/dropping probability and throughput are obtained assuming abundant control bandwidth, i.e., a sufficient number of signaling channels.

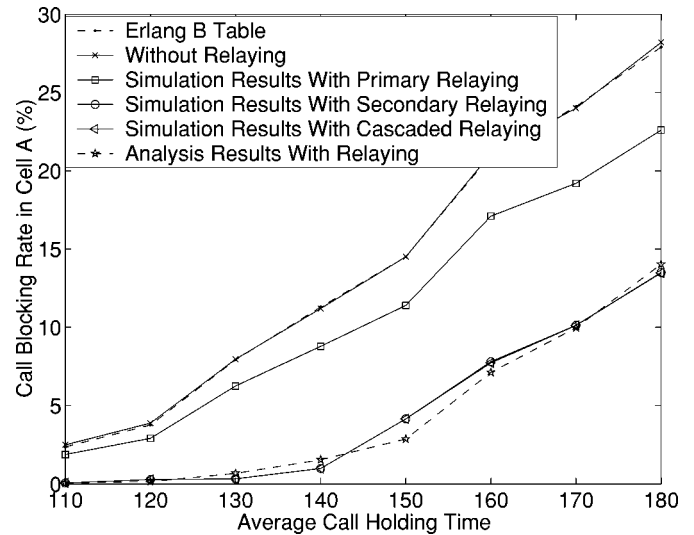


Fig. 6. Blocking probability in cell A.

bility in cell A with stationary MHs. Without any relaying, as expected, the call blocking probability which increases with traffic intensity is very close to that shown in the Erlang B table (which verifies that the simulation model is reasonable).

We observe from Fig. 6 that there is a good match between analysis and simulation results with primary and secondary relaying. Minor differences may be attributed to the fact that in the analysis we try to balance the load by relaying traffic even if there is no instantaneous blocking in that cell, whereas in simulation relaying is attempted on a call-by-call basis whenever there is blocking.

With primary relaying, the call blocking probability can be reduced but not by much. When traffic load is not very high (average holding time is less than 110 s), primary relaying can reduce the blocking probability to an acceptable level (e.g., less than 2%).

Secondary relaying reduces the call blocking probability much further. More specifically, the acceptable maximum blocking probability is normally 2%. By applying relaying, the capacity of cell A can increase from 40.255 Erlang (with holding time of 110 s) to 51.816 Erlang (with holding time of more than 140 s), which implies that the cell can take several hundred additional calls per hour and still keep the blocking probability below 2%.

Our simulation also reveals that among over 13 000 calls generated in cell A, no more than ten of them can successfully establish a cascaded relaying route. This is because after primary and secondary relaying, most of the ARSs in cell A and tier B cells have already been used to relay calls from cell A to B_i and from B_i to C_j respectively, and the active MHs using a DCH in cell A and B_i are most likely not covered by an ARS; hence either one cannot find an active MH in cell A for a secondary relaying from A to B (as the first step in cascaded relaying), or even if such an MH is found in cell A, one cannot find an active MH in cell B to complete the cascaded relaying. This is why the curves for cascaded relaying in Fig. 6 (and all following figures) almost overlap with that for secondary relaying, implying that the cascaded relaying is not very helpful.

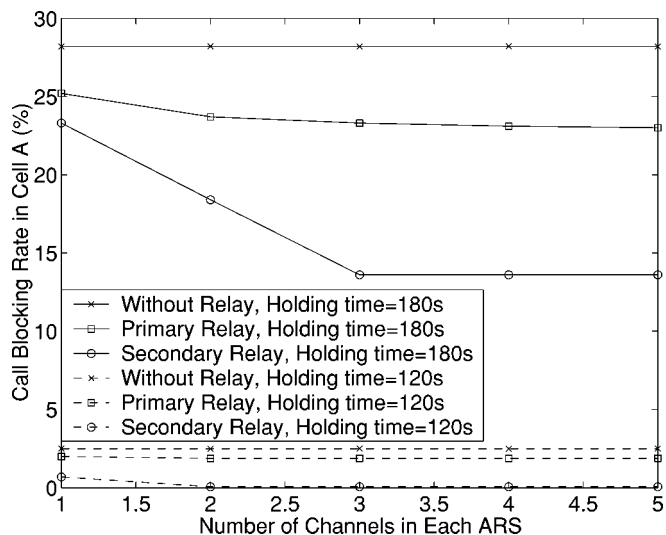


Fig. 7. Blocking probability versus number of relaying channels in cell A.

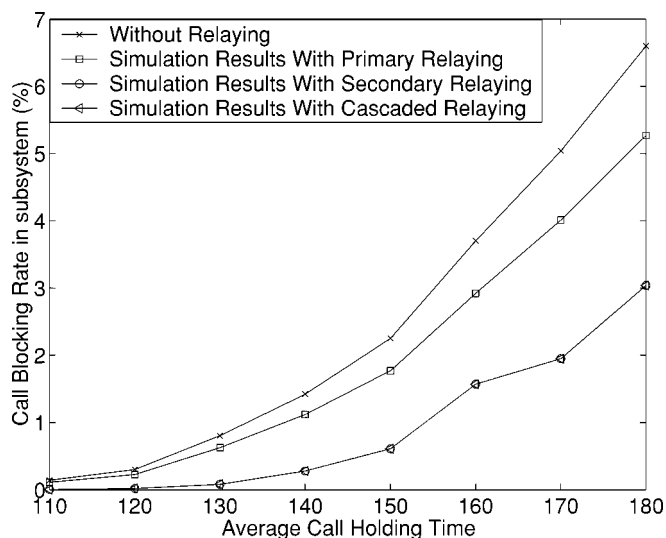


Fig. 8. Blocking probability in the entire subsystem.

Fig. 7 shows the impact of the relaying bandwidth (i.e. the number of cellular band channels each ARS can handle) on the performance. Although a higher traffic intensity may require more relaying bandwidth in order to achieve the lowest possible blocking probability in cell A, at most three cellular band channels need to be handled by each ARS for relaying purposes. Since cell A is the most congested cell (which needs to relay the largest amount of traffic), this number of channels is also enough for ARSs in cell Bs and Cs. This explains why the analytical results (which are based on the assumption that an ARS can handle as many cellular band channel as necessary) agree so well with the simulation results.

Fig. 8 shows the blocking probability of the entire subsystem. It is much lower than the results in cell A because all other cells have lower load than A. As one can see from the figure, the results due to relaying are fairly good. In particular, the system-wide blocking probability decreases although the blocking probability in other low-load cells may increase slightly because of the extra traffic relayed from the hot spot cell A. This agrees with Theorem 1 and Theorem 2 presented

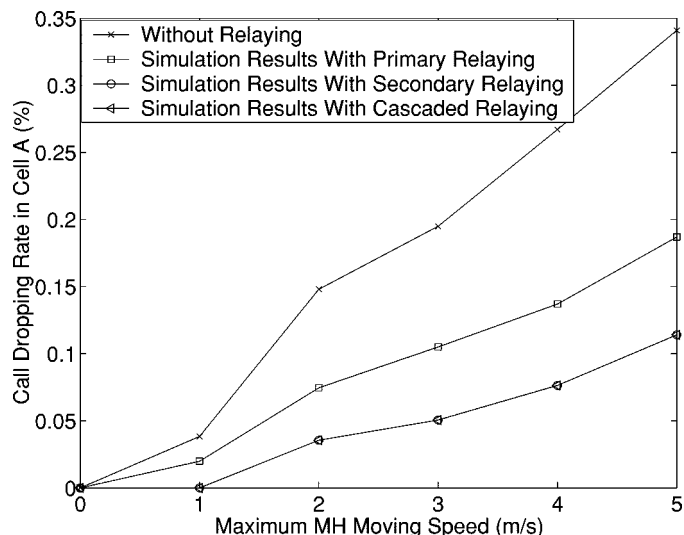


Fig. 9. Dropping probability in cell A with average holding time = 120 s.

in Section III, which prove that the iCAR system has the lowest blocking probability. Similar to the results in cell A, secondary relaying significantly reduces the call blocking probability, but cascaded relaying is only marginally useful. Though the results are not shown, mobility does not have any significant effect on the blocking probability in cell A or in the subsystem.

B. Call Dropping Probability and Handoff Performance

A call may be dropped when the active MH moves into a congested cell. In this simulation, we assume that there are no DCHs reserved for handoff calls, i.e., the handoff calls have no special priority [20]. Fig. 9 shows the dropping probability versus the maximum MH moving speed. With a higher MH mobility, the dropping probability increases sharply (recall that this is not the case for the blocking probability). In addition, when comparing with Fig. 6, we see that primary relaying performs very well for handoff calls. For example, only about 20% blocked calls are saved by primary relaying. But for handoff calls, the primary relaying can reduce the dropping probability as much as 50%. There are two reasons for the good performance of primary relaying in handoffs. First, when a call is handed off from cell X to cell Y (which is congested), it is almost guaranteed that cell X has at least one free DCH (which is released by this MH). Second, handoffs always happen at boundaries of cells, where we put the ARSs. Since a cell is modeled as a hexagon, from Table I(a), we can see that a large portion of the boundaries of a cell is covered by the ARSs. In addition, secondary relaying reduces the dropping probability further to a certain level. But due to similar reasons to those mentioned in the previous subsection, cascaded relaying is not more helpful than secondary relaying.

C. Throughput

In our simulation, we assume the transmission and reception buffer size to be zero. In other words, if a call is blocked or dropped, all the packets to be transmitted will be discarded immediately. We compare the throughput of the iCAR system with that of a cellular system (without relaying) by computing the

TABLE I
 (a) DEFAULT SIMULATION PARAMETERS. (b) MAPPING FROM AVERAGE HOLDING TIME TO TRAFFIC INTENSITY IN CELL A WITH NO MOBILITY AND EVENLY DISTRIBUTED MHs

Cell Radius (R)	2 Km
Cell Number	25
ARS Radius (r)	500 m
ARS Number	56
MH Number	25600
Simulation Area	12Km x 15.6Km
DCH at each BTS	50
DCH at each ARS	3
Average MH Call Generation Rate	1 per hour

(a)

Average Holding Time (s)	Traffic Intensity (Erlangs)
110	40.9
120	43.0
130	47.6
140	50.7
150	53.6
160	59.4
170	62.4
180	66.3

(b)

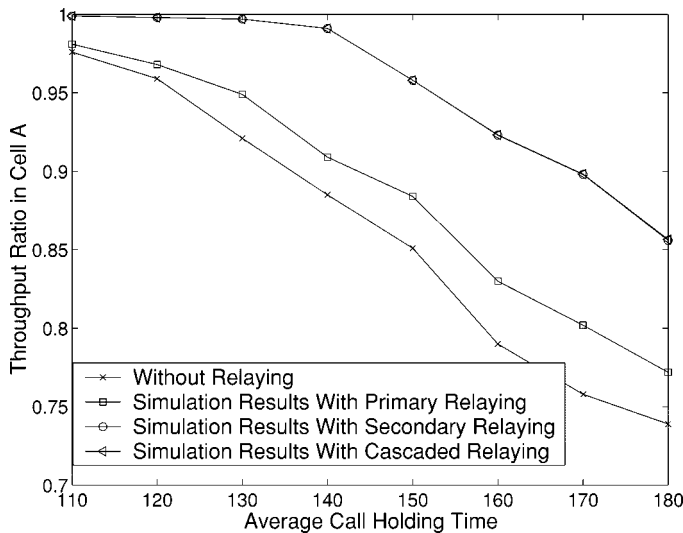


Fig. 10. Throughput in cell A.

throughput ratio, which is defined to be the ratio of received data over the data to be transmitted. This ratio is inversely proportional to the blocking/dropping probability. Fig. 10 shows the results in cell A. In general, a higher traffic load results in a lower throughput ratio because of the limited capacity. When the traffic load in cell A is low enough (with an average holding time of less than 140 seconds), we can obtain above 99% throughput ratio by applying relaying. Under a higher traffic load, one can still improve the throughput by as much as 15%. For reasons similar to those discussed in Section IV-B, cascaded relaying results in minor performance improvement. Though the results are not shown, we note that, for the overall subsystem, one can keep the throughput ratio as high as about 97%. Furthermore, with a higher MH moving speed, throughput decreases but not as dramatically as the increase in the dropping probability with the MH moving speed. This is because most of the packets are discarded during call blocking or in other words, the blocking probability dominates the throughput performance.

D. Signaling Overhead

An undesired side effect due to relaying is the extra signaling overhead. In addition to ARSs, three system components, MSC,

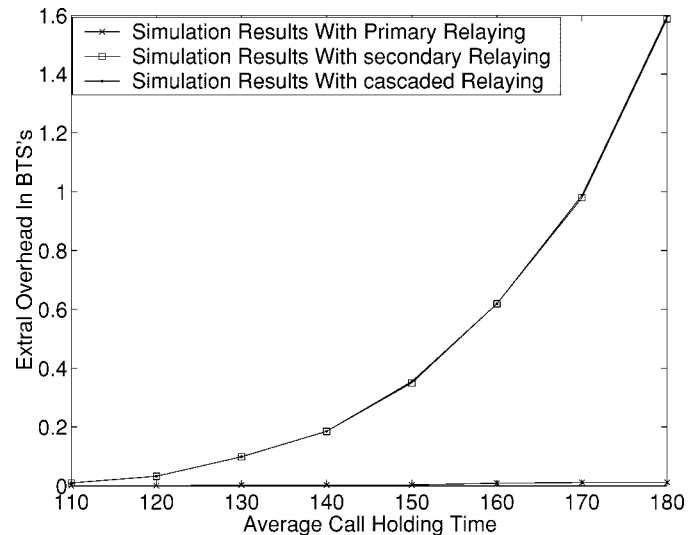


Fig. 11. Extra overhead incurred by BTSs in the subsystem.

BTS, and MH, have to send and receive more signaling packets than the case without relaying. In simulation, we study the ratio of additional amount of signaling traffic due to primary, secondary, and cascaded relaying over the basic amount of signaling traffic without relaying.

A simple signaling protocol described in [2] is implemented in the simulator. Our results (though not shown) indicate that the relaying does not add much burden to MSC. More specifically, primary relaying results in only 1% more overhead. Even in the case when one applies all three kinds of relaying, the additional overhead is at most 20%. This is reasonable because MSC does not get involved much in relaying operations.

Fig. 11 shows the extra signaling overhead incurred at a BTS when the maximum MH moving speed is 1 m/s. As can be seen from the figure, primary relaying does not cause much overhead. But when the traffic load in the system is very heavy, BTSs experience significantly high overhead while using secondary and cascaded relaying. This is because with increase in the traffic load, the probability that a call needs to be relayed also increases. This results in a large number of requests for secondary relaying. For each request, the BTS will query MSC for DCH status information, send a broadcast message to all MHs (for secondary relaying), and process replies from the MHs.

Our results also showed that the MHs suffer a higher overhead (as much as 2.5 times more than the case without relaying). This is because whenever a call tries secondary or cascaded relaying, all the active MHs using DCH in the cell are responsible for processing and replying to the broadcast messages from BTS.

Notice that the high overhead in BTSs and MHs is incurred only under very heavy traffic load (which may be unreasonably high because the blocking probability would be much more than 2%) and based on nonoptimal signaling protocols. With a normal traffic intensity with average holding time equal or less than 120 s in this simulation, the extra overhead introduced by using all three kinds of relaying at MSC, BTSs, and MHs are only about 1%, 3%, and 5%, respectively, which is not significant. Nevertheless, further research is needed to improve the signaling protocols to reduce the overhead and to study the tradeoff introduced by dedicating one or more additional channels to carry control signaling information.

Finally, our simulation results also revealed that although with a higher MH moving speed, the MHs need to process more signaling messages because of the higher probability that a handoff call needs relaying in order to avoid being dropped, mobility has little effect on the signaling overhead.

V. RELATED WORK

In this section, we discuss a few related studies in the literature. In [21], the authors presented a hierarchical structure for wireless mobile systems with a fixed backbone. In order to access the backbone, all MHs have to go through a mobile base station (which can be thought of as a cluster head). It is similar to iCAR in that the cellular infrastructure in iCAR is also fixed, and the ARSs can be mobile and used to relay between MHs and the fixed BTSs. However, in iCAR, the MHs have two air (or radio) interfaces so that they may communicate with BTSs directly without going through ARSs. In addition, each ARS is under the control of a MSC and has limited mobility. Such a feature is important to ensure that a relaying route can be set up fast and maintained with a high degree of stability. Routing in iCAR is similar to that of having a hybrid (both hierarchical and flat) structure in [22] for efficient routing and handoffs in mobile ATM networks. The difference between the two is that in the latter, path extension (or relaying) is between two (fixed) BTSs through direct wired links.

In the multihop cellular systems approach [23] and the mobile-assisted connection admission (MACA) system [24], relaying is performed by MHs, and thus that approach shares many disadvantages in terms of security (authentication, privacy), billing, and mobility management (of the MHs) with mobile *ad hoc* networks. In addition, the main goal of the multihop cellular systems is to reduce the number of BTSs or the transmission power of each BTS, but it can no longer guarantee a full coverage of the area. In fact, even in the ideal case where every MH in an area uncovered by any BTS can find a relaying route (through other MHs), the multihop approach will neither increase the system capacity nor decrease the call blocking/dropping probability, unless a large percentage of the calls are intracell calls (i.e., calls whose source and destination are in the same cell), which usually is not the case in practice.

Note that the proposed relaying through ARSs is useful in any cellular system where congestion may occur, even though a call may not be allocated a dedicated DCH all the time (or in other words, during the entire call duration). Also, if one simply treats the 2.4-GHz band as an additional set of channels that can be used in a cellular system (by, e.g., modifying each BTS so it is equipped with the R-interface as well), one will not be able to balance loads among cells or to eliminate congestion in hot-spot cells via relaying. Other approaches such as those using microwave links between BTSs, cell splitting, cell sectorization, and cell breathing cannot serve as a replacement for relaying in iCAR either, although they may be used in conjunction with our approach.

VI. CONCLUSION

We have proposed a novel architecture for next-generation wireless systems called iCAR which integrates the traditional cellular and modern relaying technologies. We have also evaluated the performance improvement of iCAR over conventional cellular systems under Erlang-B traffic model. The basic idea of the iCAR is to place a number of ARSs in a cellular system to divert excess traffic from one (possibly congested) cell to another. We have compared the performance of the iCAR system with the conventional cellular system via analysis and simulations in terms of the call blocking/dropping probability, throughput, and signaling overhead in both the hot cells and overall subsystem. Our results have shown that iCAR, with only a limited number of ARSs placed using the seed-growing approach (see the end of Section II), can dynamically balance the traffic among cells, reduce the call blocking/dropping probability (thus increase system capacity), and improve the system throughput cost effectively.

APPENDIX A

Proof of Theorem 1

Theorem 1: Assume that the total traffic in an n -cell system is T Erlangs, then the (system wide) call-blocking probability is minimized when the traffic in each cell is T/n Erlangs.

Proof: Let the number of DCHs in each cell be M and assume that the traffic intensity is T_i in each cell i where $T = \sum_{i=1}^n T_i$. The probability of all the channels in cell i being busy is given by the following Erlang B formula:

$$B_i(M; T_i) = \frac{T_i^M / M!}{\sum_{i=0}^M T_i^i / i!}.$$

For the n -cell system, the average blocking probability for the entire system is

$$B = \frac{\sum_{i=1}^n B_i \times T_i}{T}.$$

Since $T = \sum_{i=1}^n T_i$, we may write $T_n = T - \sum_{i=1}^{n-1} T_i$. In other words, there are only $n - 1$ independent T_i s. In order to

compute the minimum value of B , we compute all the partial derivatives of B and set them to be zero, that is

$$\frac{\partial B}{\partial T_i} = 0 \quad (1 \leq i \leq n-1).$$

We omit the details but we can obtain the critical points (or the solutions to the above equations) as

$$T_1 = T_2 = \dots = T_n = \frac{T}{n} = \bar{T}.$$

By computing the second-order partial derivatives of B which forms a matrix, and by verifying that its determinant is larger than zero at the above critical points, we have shown that the blocking probability reaches its minimum value when the traffic is evenly distributed (i.e., $T_i = T_j$ for any $1 \leq i, j \leq n$ and $i \neq j$). ■

APPENDIX B

Proof of Theorem 2

Theorem 2: For a given total traffic intensity and a fixed number of DCHs in each cell, an ideal iCAR has a lower blocking probability than any conventional cellular systems (including a perfectly load-balanced one).

Proof: Given that the iCAR system may be treated as a super cell with a total of $T = n\bar{T}$ Erlangs and nM channels (where n is the number of cells in the system, \bar{T} and M are the average traffic intensity and the number of DCHs in each cell, respectively), the blocking probability is

$$B(nM; n\bar{T}) = \frac{(n\bar{T})^{nM} / (nM)!}{\sum_{i=0}^{nM} (n\bar{T})^i / i!}.$$

According to Theorem 1, the minimum blocking probability of any conventional cellular system with \bar{T} Erlangs and M DCHs in each cell is

$$B(M; \bar{T}) = \frac{(\bar{T})^M / (M)!}{\sum_{i=0}^M (\bar{T})^i / i!}.$$

We prove that $B(nM; nX) < B(M; X)$ for any $n, X > 1$ by showing that $1/B(nM; nX) > 1/B(M; X)$ as follows:

$$\begin{aligned} \frac{1}{B(M; X)} &= \frac{M! \sum_{i=0}^M X^i / i!}{X^M} \\ &= \sum_{i=0}^M \frac{M!}{X^{M-i} i!} \\ \text{Let } j &\equiv M-i \\ &\sum_{j=0}^M \frac{M(M-1) \dots (M-j+1)}{X^j}. \end{aligned}$$

Similarly

$$\begin{aligned} \frac{1}{B(nM; nX)} &= \sum_{j=0}^{nM} \frac{nM(nM-1) \dots (nM-j+1)}{(nX)^j} \\ &= \sum_{j=0}^{nM} \frac{M \left(M - \frac{1}{n} \right) \dots \left(M - \frac{j+1}{n} \right)}{X^j}. \end{aligned}$$

Since every term in the above equation is positive

$$\begin{aligned} \frac{1}{B(nM; nX)} &> \sum_{j=0}^M \frac{M \left(M - \frac{1}{n} \right) \dots \left(M - \frac{j+1}{n} \right)}{X^j} \\ &> \sum_{j=0}^M \frac{M(M-1) \dots (M-j+1)}{X^j} \\ &= \frac{1}{B(M; X)}. \end{aligned} \quad \blacksquare$$

REFERENCES

- [1] C. Qiao, H. Wu, and O. Tonguz, "Load balancing via relay in next generation wireless systems," in *Proc. IEEE Conf. Mobile Ad Hoc Networking Computing*, Aug. 2000, pp. 149–150.
- [2] C. Qiao and H. Wu, "iCAR: An integrated cellular and ad-hoc relay system," in *IEEE Int. Conf. Computer Communication Network*, Oct. 2000, pp. 154–161.
- [3] [Online]. Available: <http://www.nwr.nokia.com/>.
- [4] T. Rappaport, *Wireless Communications Principle and Practice*. New York: Prentice-Hall, 1996.
- [5] V. Garg and J. Wilkes, *Wireless and Personal Communications Systems*. New York: Prentice-Hall, 1996.
- [6] S. Das, R. Castaneda, J. Yan, and R. Sengupta, "Comparative performance evaluation of routing protocols for mobile, ad hoc networks," in *7th Int. Conf. Computer Communications Networks (IC3N)*, 1998, pp. 153–161.
- [7] Y.-B. Ko and N. H. Vaidya, "Location-aided routing (LAR) in mobile ad hoc networks," in *ACM/IEEE 4th Ann. Int. Conf. Mobile Computing Networking (MobiCom 98)*, 1998.
- [8] C. Toh, *Wireless ATM and Ad-Hoc Networks: Protocols and Architectures*. New York: Kluwer, 1996.
- [9] C. Perkins and P. Bhagwat, "Highly dynamic destination sequenced distance vector routing (dsv) for mobile computers," in *Proc. ACM SIGCOMM '94*, 1994, pp. 234–244.
- [10] C. Perkins and E. Royer, "Ad-hoc on demand distance vector routing," in *Proc. IEEE WMCSA '99*, 1999, pp. 90–100.
- [11] V. Park and M. Corson, "A highly adaptive distributed routing algorithm for mobile wireless networks," in *Proc. IEEE INFOCOM '97*, 1997, pp. 1405–1413.
- [12] D. Johnson and D. Maltz, "Dynamic source routing in ad hoc wireless networks," *Mobile Computing*, vol. 5, pp. 153–181, 1996.
- [13] S. Murthy and J. Garcia-Luna-Aceves, "An efficient routing protocol for wireless networks," *ACM/Baltzer Mobile Networks Applicat.*, vol. 1, no. 2, pp. 183–197, 1996.
- [14] A. Iwata, C.-C. Chiang, G. Pei, M. Gerla, and T.-W. Chen, "Scalable routing strategies for ad-hoc wireless networks," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1369–1379, 1999.
- [15] S. Basagni, I. Chlamtac, V. Syrotiuk, and B. Woodward, "A distance routing effect algorithm for mobility (dream)," in *Proc. ACM/IEEE MobiCom '98 Conf.*, 1998, pp. 76–84.
- [16] G. Stuber, *Principles of Mobile Communication*. New York: Kluwer, 1996.
- [17] R. Kohno, R. Meidan, and L. Milstein, "Spread spectrum access methods for wireless communications," *IEEE Commun. Mag.*, vol. 33, pp. 58–67, 1995.
- [18] X. Zeng, R. Bagrodia, and M. Gerla, "GloMoSim: A library for parallel simulation of large-scale wireless networks," in *Proc. Workshop Parallel and Distributed Simulation*, 1998, pp. 154–161.
- [19] H. Jung and O. K. Tonguz, "Random spacing channel assignment to reduce the nonlinear intermodulation distortion in cellular mobile communications," *IEEE Trans. Veh. Technol.*, vol. 48, pp. 1666–1675, 1999.
- [20] H. Ebersman and O. K. Tonguz, "Handoff ordering using signal prediction priority queueing in personal communication systems," *IEEE Trans. Veh. Technol.*, vol. 48, pp. 20–35, 1999.
- [21] I. F. Akyildiz, W. Yen, and B. Yener, "A new hierarchical routing protocol for dynamic multihop wireless networks," in *IEEE INFOCOM '97*, 1997, pp. 1422–1429.
- [22] S. Maloo and C. Qiao, "Efficient routing and fast handoff in a mobile atm network with a hybrid topology," in *5th Int. Conf. Info. Systems Analysis Synthesis (ISAS)*, vol. 4, 1999, pp. 614–621.
- [23] Y. D. Lin and Y. C. Hsu, "Multihop cellular: A new architecture for wireless communication," in *IEEE INFOCOM '2000*, 2000, pp. 1273–1282.

- [24] X. Wu, B. Mukherjee, and G. Chan, "MACA—An efficient channel allocation scheme in cellular networks," in *IEEE GLOBECOM*, Dec. 2000, pp. 1385–1389.



Hongyi Wu received the B.S. degree in electrical engineering from Zhejiang University, China, in 1996 and the M.S. degree in electrical engineering from State University of New York at Buffalo in February 2000. He is currently a Ph.D. candidate in the Department of Computer Science and Engineering in State University of New York at Buffalo.

He worked at the Nokia Research Center as an Intern in the summers of 2000 and 2001. His research interests include wireless mobile *ad hoc* networks, cellular systems, 3G system, routing protocols, and

the Internet.



Chunming Qiao received the B.S. degree in computer science from University of Science and Technology of China. He received the Ph.D. in computer science from the University of Pittsburgh in 1993.

He is currently a tenured Associate Professor at the Computer and Science Engineering Department. He is also an Adjunct Professor at the Electrical Engineering Department, and the Director of the Laboratory for Advanced Network Design, Evaluation and Research (LANDER), which conducts cutting-edge research related to optical networks, wireless/mobile

networks, and the Internet. He has over ten years of research experience in optical networks, covering the areas of photonic switching devices, WDM communications, and optical packet-switching. He pioneered research on the next-generation Optical Internet, and in particular, optical burst switching (OBS). He has published more than six dozen papers in leading technical journals and conferences, and given several keynote speeches, tutorials, and invited talks.

Dr. Qiao is the Chair of the program section on IP over WDM at the upcoming first Asia-Pacific Optical Wireless Communications (APOC) 2001. He has served as the Chair of the program on Optical Layer and Internetworking Technology in 2000, a Co-Chair for the annual All-Optical Networking Conference from 1997 to 2000, a Program Vice Co-Chair for the 1998 International Conference on Computer Communications and Networks (IC3N), an expert panelist, a technical program committee member, and session organizer/chair in several other conferences and workshops. He is also the Founder and Chair of the recently established Technical Group on Optical Networks (TGON) sponsored by SPIE. He is the IEEE Communication Society's Editor-at-Large for optical networking and computing, an editor of *IEEE/ACM TRANSACTIONS ON NETWORKING*, the *Journal on High-Speed Networks*, and the *Optical Networks Magazine*. He received the Andrew-Mellon Distinguished doctoral fellowship award from the University of Pittsburgh.



Swades De received the B.Tech. degree in radio-physics and electronics from University of Calcutta in 1993 and the M.Tech. degree in optoelectronics and optical communication from the Indian Institute of Technology Delhi in 1998. He is a Ph.D. candidate in the Electrical Engineering Department at State University of New York at Buffalo.

From 1993 to 1997 and in 1999, he worked in different telecommunication companies in India as a Hardware and Software Development Engineer. His current research interests include performance study, resource optimization in high speed networks,

routing in multihop wireless networks, load balancing in cellular wireless networks, and communications and systems issues in optical networks.



Ozan Tonguz received the B.Sc. degree from the University of Essex, England, in 1980 and the M.Sc. and Ph.D. degrees from Rutgers University, New Brunswick, NJ, in 1986 and 1990, respectively, all in electrical engineering.

He is currently a tenured Full Professor in the Department of Electrical and Computer Engineering of Carnegie Mellon University (CMU). Before joining CMU, he was on the faculty of the ECE Department of the State University of New York (SUNY). He joined SUNY/Buffalo in 1990 as an

Assistant Professor, was granted early tenure, and was promoted to Associate Professor in 1995 and to Full Professor in 1998. Prior to joining academia, he was with Bell Communications Research (Bellcore) between 1988 and 1990 doing research in optical networks and communication systems. His current research interests include optical networks, wireless networks and communication systems, high-speed networking, and satellite communications. The author or coauthor of more than 100 technical papers in IEEE journals and conference proceedings, and a book chapter (Wiley, 1999), his contributions in optical networks and communication systems and wireless networks are internationally acclaimed. His industrial experience includes periods with Bell Communications Research, CTI, Inc., Harris RF Communications, Nokia Networks, Aria Wireless Systems, and Clearwire Technologies. He currently serves as a consultant to several industrial and government organizations in the U.S. and Europe in the broad area of telecommunications (optical networks, wireless communications, and high-speed networking). He is also a Co-Director (Thrust Leader) of the new Center for Wireless and Broadband Telecommunications at Carnegie Mellon University.

Dr. Tonguz serves or has served as an Associate Editor for the *IEEE TRANSACTIONS ON COMMUNICATIONS*, *IEEE Communications Magazine*, and *JOURNAL OF LIGHTWAVE TECHNOLOGY*. He was a Guest Editor of special issues of the *JOURNAL OF LIGHTWAVE TECHNOLOGY* and *IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS*.